Ⅵ 통계

자료의 특징을 알려면

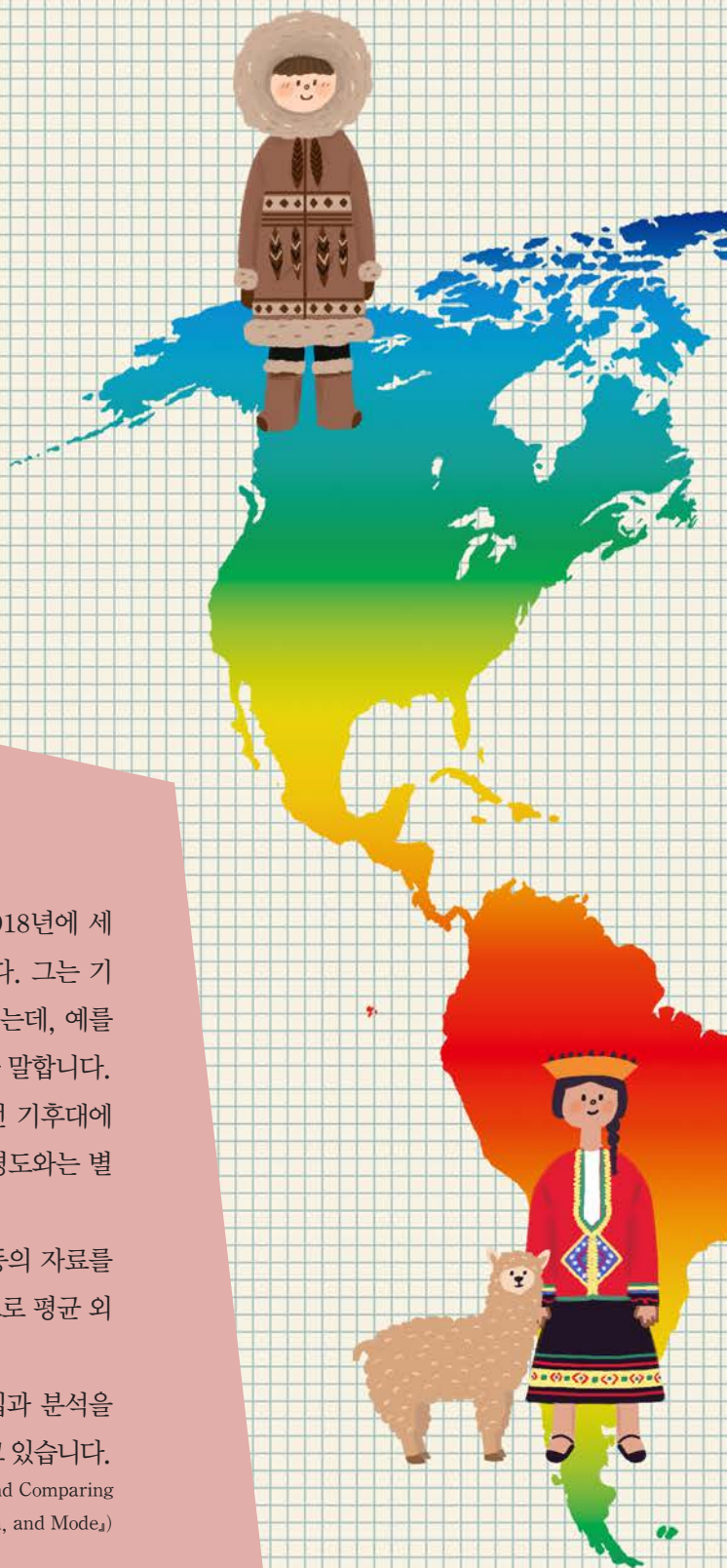
독일의 기후학자 쾨펜(Köppen, W., 1846~1940)은 1918년에 세계의 기후대를 열대, 건조대, 온대, 냉대, 한대로 구분했습니다. 그는 기후대를 기온과 강수량의 월간 및 연간 평균에 근거하여 구분했는데, 예를 들어 열대는 가장 추운 달의 평균 기온이 18 ℃ 이상인 지역을 말합니다. 쾨펜이 구분한 기후대로부터 알 수 있듯이, 어떤 지역이 어떤 기후대에 속하는지는 그 지역의 위도와는 매우 밀접한 관계가 있지만 경도와는 별다른 관계가 없습니다.

기후학은 광범위한 지역의 기온, 강수량, 적설량, 일조량 등의 자료를 많이 다루는데, 이러한 자료들의 경향과 특징을 대표하는 값으로 평균 외에 다른 값들도 사용한다고 합니다.

한편, 컴퓨터 성능의 획기적 발전에 힘입어 빅 데이터 수집과 분석을 통해 통계는 이전보다 더 많은 부분에서 우리 삶에 영향을 끼치고 있습니다.

(출처: Linde, B. M., 『Climates of the World: Identifying and Comparing Mean, Median, and Mode』)

이 단원에서는

자료의 경향을 대표하는 여러 가지 대푯값의 의미를 이해하고, 자료에서 두 변량 사이의 상관관계를 배웁니다.

1 대푯값과 산포도

평균 키나 평균 점수와 같이 평균은 어떤 자료의 경향을 대표하는 값으로 우리가 가장 흔히 사용하는 통계적 개념입니다.

그런데 유치원 어린이 5명과 키다리 아저씨 1명이 있을 때, 이 6명의 키의 평균보다는 키의 크기를 순서대로 나열하여 중앙에 오는 값이 오히려 키에 대한 대표적 경향을 더 잘 나타낸다고 말할 수 있습니다.

한편, 고대 아테네 사람들은 멀리 보이는 성벽의 높이를 짐작하기 위해서 많은 사람들에게 그 성벽의 벽돌이 수직으로 몇 층 쌓여있는지 각각 헤아리게 했다고 합니다. 다수의 사람들이 벽돌 수를 제대로 세었을 것으로 생각하여 그 층수에 벽돌 한 장의 높이를 곱하여 그 성벽 전체 높이를 짐작하려 했던 것입니다. 이와 같이 때로는 자료의 변량 중에서 가장 빈번하게 나타나는 값을 대표로 정하기도 합니다.

(출처: 박영희, 「통계 영역에서 대푯값의 의미와 지도에 관한 고찰」)

이 단원에서는 자료의 특징을 나타내는 대푯값과 자료의 변량들이 흩어진 정도를 나타내는 산포도에 대하여 알아봅니다.

준비 학습

• 자료의 정리와 평균

1 다음은 퀴즈 대회에 참가한 학생 10명이 맞힌 문제의 개수를 조사하여 나타낸 것이다.

맞힌 문제의 개수 (단위: 개)

7	8	9	10	10	6	5	7	10	6

(1) 맞힌 문제의 개수가 적은 쪽에서 6번째에 해당하는 변량을 구하시오.

(2) 맞힌 문제의 개수의 평균을 구하시오.

대푯값

학습 목표 • 중앙값, 최빈값, 평균의 의미를 이해하고, 이를 구할 수 있다.

중앙값이란 무엇인가?

다음은 농구 경기에서 선수 7명이 얻은 점수를 조사하여 나타낸 것이다.

선수들이 얻은 점수 (단위: 점)

3	4	7	5	8	30	6

1. 선수들이 얻은 점수의 평균을 구해 보자.

2. 점수를 작은 값부터 순서대로 나열할 때, 중앙에 위치하는 값을 구해 보자.

자료의 중심적인 경향이나 특징을 대표적으로 나타내는 값을 그 자료의 **대푯값**이라고 한다. 대푯값으로 가장 많이 쓰이는 것은 평균이다.

위의 생각 열기에서 선수들이 얻은 점수의 평균은

$$\frac{3+4+7+5+8+30+6}{7}=9(\text{점})$$

이다. 그런데 이 자료에서 선수 6명이 얻은 점수는 평균보다 낮고, 1명이 얻은 점수는 평균보다 훨씬 높으므로 평균 9점은 이 자료의 중심적인 경향을 잘 나타낸다고 할 수 없다.

이와 같이 자료의 변량 중에 매우 크거나 작은 값이 포함되어 있는 경우에는 평균이 그 값에 영향을 많이 받기 때문에 평균 이외의 다른 대푯값을 생각할 필요가 있다.

배웠어요!
변량은 자료를 수량으로 나타낸 것이다.

자료의 변량을 작은 값부터 순서대로 나열할 때, 중앙에 위치하는 값을 그 자료의 **중앙값**이라고 한다.

이를테면 생각 열기의 자료의 변량을 작은 값부터 순서대로 나열하면

3, 4, 5, 6, 7, 8, 30

이므로 중앙값은 6이다.

일반적으로 자료의 변량 중에서 매우 크거나 작은 값이 포함되어 있는 경우에는 중앙값이 평균보다 그 자료의 중심적인 경향을 더 잘 나타낼 수 있다.

자료의 변량을 작은 값부터 순서대로 나열할 때, 변량의 개수가 홀수이면 가운데 위치하는 값을 중앙값으로 하고, 변량의 개수가 짝수이면 가운데 위치하는 두 값의 평균을 중앙값으로 한다.

보기 ① 자료 '9, 8, 6, 3, 9'의 변량을 작은 값부터 순서대로 나열하면

3, 6, 8, 9, 9

이고 변량이 5개이므로, 이 자료의 중앙값은 세 번째 값인 8이다.

② 자료 '13, 15, 17, 21, 26, 31'의 변량은 6개이므로, 이 자료의 중앙값은 세 번째 값과 네 번째 값의 평균인 $\frac{17+21}{2}=19$이다.

문제 1 다음 자료의 중앙값을 구하시오.

(1) 12, 21, 17, 21, 19, 14, 33　　(2) 3, 6, 26, 133, 49, 14, 28, 40

탐구 문제 2 다음은 2015년 APEC 회원국 중에서 15개 국가의 인구 밀도를 조사하여 나타낸 것이다.

▸ APEC은 '아시아태평양경제협력체'로 2018년 현재 가입 국가는 총 21개국이다.

인구 밀도　(단위: 명/km²)

국가	인구 밀도	국가	인구 밀도	국가	인구 밀도
뉴질랜드	17	브루나이	72	중국	143
대한민국	509	싱가포르	7698	칠레	24
말레이시아	94	오스트레일리아	3	캐나다	4
멕시코	62	인도네시아	134	페루	24
미국	33	일본	336	필리핀	339

(출처: 국가통계포털, 2018)

(1) 인구 밀도의 평균과 중앙값을 각각 구하시오.

(2) 평균과 중앙값 중에서 인구 밀도의 대푯값으로 적절한 것을 선택하고, 대푯값으로 선택한 이유를 말하시오.

최빈값이란 무엇인가?

생각 열기 다음 줄기와 잎 그림은 어느 신발 가게에서 하루 동안 판매한 운동화의 크기를 조사하여 나타낸 것이다.

판매한 운동화의 크기 (23|5는 235 mm)

줄기	잎
23	5
24	0 0 5 5
25	0 0 5
26	0 0 0 0 5
27	0 5

1. 판매한 운동화의 크기의 평균과 중앙값을 구해 보자.

2. 가장 많이 판매한 운동화의 크기를 말해 보자.

위의 생각 열기에서 판매한 운동화의 크기의 평균은 254 mm이고 중앙값은 255 mm이지만, 가장 많이 판매한 운동화의 크기는 260 mm이다.

이때 이 신발 가게에서는 평균이나 중앙값보다 가장 많이 판매한 운동화의 크기가 더 유용하게 쓰일 수 있다.

▸ 최빈값에서 최빈(最頻)은 '가장 자주'라는 뜻이다.

이와 같이 자료의 변량 중에서 가장 많이 나타나는 값을 그 자료의 **최빈값**이라고 한다.

일반적으로 최빈값은 변량이 중복되어 나타나는 자료나, 가장 좋아하는 운동 종목처럼 숫자로 나타낼 수 없는 자료의 대푯값으로 유용하다.

한편, 값이 하나로 정해지는 평균이나 중앙값과는 달리 최빈값은 자료에 따라 두 개 이상일 수도 있다.

보기 ① 자료 '2, 5, 7, 7, 1, 7, 5'에서 7이 세 번으로 가장 많이 나타나므로, 이 자료의 최빈값은 7이다.

② 자료 '사과, 오렌지, 바나나, 사과, 파인애플, 사과'에서 사과가 가장 많이 나타나므로, 이 자료의 최빈값은 사과이다.

③ 자료 '5, 8, 7, 2, 7, 9, 8'에서 7과 8이 각각 두 번씩 가장 많이 나타나므로, 이 자료의 최빈값은 7과 8이다.

문제 3 다음 자료의 최빈값을 구하시오.

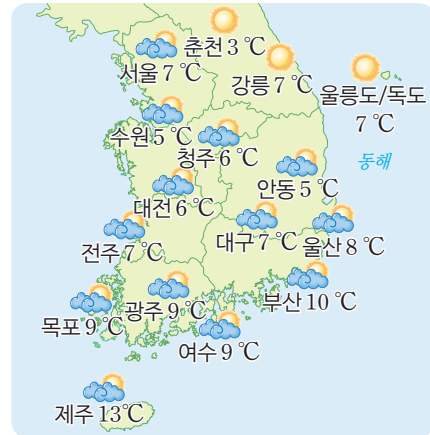
(1) 2, 5, 5, 8, 3, 6, 5, 8, 4

(2) 35, 28, 71, 28, 37, 35, 48

(3) 빨강, 빨강, 노랑, 파랑, 빨강, 파랑

문제 4 오른쪽 그림은 어느 날 오전 8시 우리나라 여러 지역의 기온을 조사하여 나타낸 것이다.

(1) 기온의 중앙값을 구하시오.

(2) 기온의 최빈값을 구하시오.

대푯값으로 평균이 가장 많이 쓰이지만 자료에 따라 중앙값이나 최빈값이 자료의 특징을 더 잘 나타낼 수도 있으므로, 자료의 특성에 따라 적절한 대푯값을 선택하는 것이 중요하다.

▶ 평균, 중앙값, 최빈값 중에서 다음 각 자료의 특징을 더 잘 나타낼 수 있는 대푯값을 선택하여 구하고, 대푯값으로 선택한 이유를 말해 보자.

(1) 중학생 10명이 한 달 동안 읽은 책의 권수 (단위: 권)

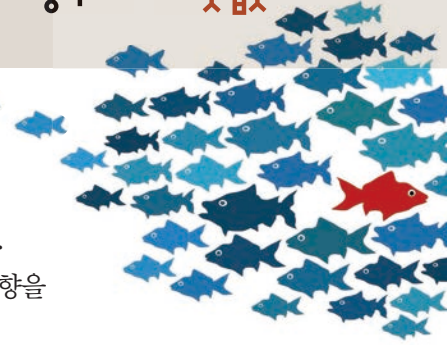
3　4　1　5　4　6　2　3　5　20

(2) 우리 반 학생 10명의 수학 점수 (단위: 점)

84　75　88　92　96　76　83　98　100　78

(3) 편의점에서 하루 동안 팔린 우유 10개의 용량 (단위: mL)

180　240　300　240　500　180　240　240　240　300

이상점이 있는 경우의 대푯값

통계학에서는 수집된 자료에서 대부분의 변량들이 갖는 성향과 동떨어진 성향을 갖는 변량을 이상점(異常點, outlier)이라고 한다. 평균은 이상점에 매우 민감한 영향을 받지만, 중앙값은 이상점에 영향을 덜 받는다.

이상점은 평균에 큰 영향을 미치므로 이상점이 있는 자료에서 평균을 이용하는 통계 분석은 타당성을 잃을 염려가 있다. 따라서 이런 경우에 몇 개의 이상점을 제외하고 평균을 구하는 방법을 사용하기도 한다.

예를 들어 19세기 프랑스의 농민들은 포도 생산량의 평균을 계산할 때 지난 20년간의 포도 생산량 중에서 가장 풍작인 해와 가장 흉작인 해를 제외한 18년 동안의 생산량의 평균을 구함으로써 이러한 문제를 해결했다고 한다.

이처럼 어떤 자료의 변량 중에서 이상점을 제외하고 평균을 구하는 것을 절사 평균(切捨平均, trimmed mean)이라고 하는데, 이것은 체조나 피겨 스케이팅 경기의 채점에 쓰인다.

(출처: 성내경, 『정보 시대 그리고 통계』)

탐구 다음은 2017년 APEC 회원국 중에서 15개 국가의 수출액을 조사하여 나타낸 것이다.

수출액 (단위: 억 달러)

국가	수출액	국가	수출액	국가	수출액
뉴질랜드	381	브루나이	56	중국	22804
대한민국	5737	싱가포르	3732	칠레	692
말레이시아	2178	오스트레일리아	2311	캐나다	4237
멕시코	4095	인도네시아	1687	페루	449
미국	15463	일본	6982	필리핀	687

(출처: 국가통계포털, 2018)

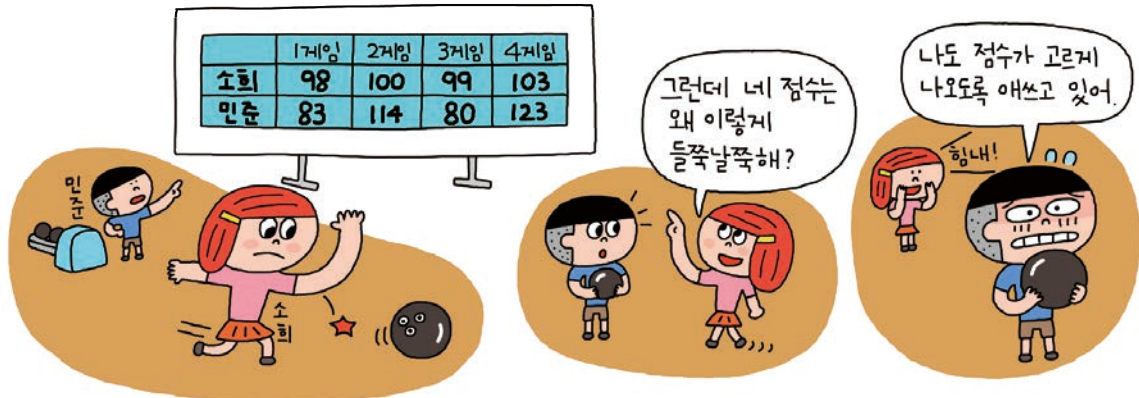
(1) 중국과 브루나이의 수출액이 이 자료의 이상점인 이유를 말해 보자.

(2) 수출액의 평균과 중앙값을 구하여 어느 것이 대푯값으로 적절한지 말해 보자.
(단, 평균은 반올림하여 일의 자리까지 구한다.)

(3) 위의 자료의 변량 중에서 최댓값과 최솟값을 제외한 13개 변량의 평균과 중앙값을 구하여 (2)에서 구한 값과 비교해 보자. (단, 평균은 반올림하여 일의 자리까지 구한다.)

산포도

학습 목표 • 분산과 표준편차의 의미를 이해하고, 이를 구할 수 있다.

다가서기

산포도란 무엇인가?

생각 열기

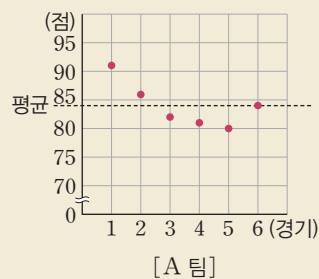
다음은 두 농구팀 A와 B가 6경기에서 얻은 점수를 조사하여 나타낸 것이다.

두 농구팀이 얻은 점수 (단위: 점)

팀 \ 경기	1	2	3	4	5	6	평균
A	91	86	82	81	80	84	84
B	75	89	80	95	73	92	84

1. A 팀이 얻은 점수를 나타낸 그래프를 보고, 같은 방법으로 B 팀이 얻은 점수를 그래프로 나타내 보자.

[A 팀]

[B 팀]

2. 평균을 중심으로 변량이 더 많이 모여 있는 팀을 말해 보자.

위의 생각 열기에서 A 팀과 B 팀이 얻은 점수의 평균은 모두 84점으로 같다. 그런데 A 팀의 점수가 B 팀의 점수에 비하여 평균을 중심으로 더 많이 모여 있음을 알 수 있다.

이와 같이 두 자료의 평균은 같지만 분포 상태는 다를 수 있다.

▸‘산포도(散布度)’는 흩어져 퍼진 정도라는 뜻이다.

자료의 분포 상태를 알아보기 위하여 변량들이 대푯값을 중심으로 흩어져 있는 정도를 하나의 수로 나타낸 값을 그 자료의 **산포도**라고 한다.

산포도에는 여러 가지가 있으나, 보통 평균을 대푯값으로 할 때의 산포도를 사용한다. 일반적으로 평균을 대푯값으로 할 때, 자료의 변량이 평균에 모여 있을수록 산포도는 작아지고, 흩어져 있을수록 산포도는 커진다.

편차, 분산과 표준편차란 무엇인가?

어떤 자료에 대하여 각 변량에서 평균을 뺀 값을 그 변량의 **편차**라고 한다.

(편차)=(변량)−(평균)

편차는 변량이 평균보다 크면 양수이고, 변량이 평균보다 작으면 음수이다. 또 편차의 절댓값이 클수록 그 변량은 평균에서 멀리 떨어져 있고, 편차의 절댓값이 작을수록 그 변량은 평균에 가까이 있다.

보기 자료 ‘1, 2, 3, 4, 5’의 평균은 3이므로 각 변량의 편차는 순서대로

$-2,\ -1,\ 0,\ 1,\ 2$

이다.

다음을 통하여 편차의 특징을 알아보자.

다음은 어느 문구점에서 하루 동안 판매한 줄넘기의 개수를 7일 동안 조사하여 나타낸 것이다. 이 자료에서 각 변량의 편차의 합을 구하려고 한다.

줄넘기의 판매량 (단위: 개)

판매량	7	11	9	6	10	8	5
편차							

1 줄넘기의 판매량의 평균을 구하고, 위의 표를 완성해 보자.

2 각 변량의 편차의 합을 구해 보자.

위의 함께하기에서 편차의 합은 0임을 알 수 있다.

일반적으로 편차의 합은 항상 0이므로 편차의 평균도 0이 되어 편차의 평균으로는 변량들이 흩어진 정도를 알 수 없다.

따라서 편차를 제곱한 값의 평균과 그 양의 제곱근을 산포도로 사용한다.

이때 편차를 제곱한 값의 평균을 **분산**이라 하고, 분산의 양의 제곱근을 **표준편차**라고 한다.

▸피어슨(Pearson, K., 1857~1936)
영국의 수학자로, '표준편차'라는 용어를 처음 사용했다.

즉, 자료의 분산과 표준편차는 다음과 같다.

분산과 표준편차

❶ $(\text{분산})=\dfrac{\{(\text{편차})^2\text{의 총합}\}}{(\text{변량의 개수})}$

❷ $(\text{표준편차})=\sqrt{(\text{분산})}$

일반적으로 자료의 분산 또는 표준편차가 클수록 그 자료의 변량들이 평균을 중심으로 흩어져 있고, 작을수록 평균을 중심으로 모여 있다.

예제 1

다음은 어느 버스 정류장을 지나는 5개의 버스 노선에 대하여 각 배차 간격을 조사하여 나타낸 것이다. 배차 간격의 분산과 표준편차를 각각 구하시오.

배차 간격 (단위: 분)

6	8	7	10	4

풀이 주어진 자료에서

$$(\text{평균})=\frac{6+8+7+10+4}{5}=7(\text{분})$$

이므로 각 변량의 편차를 구하여 표로 나타내면 다음과 같다.

배차 간격 (단위: 분)

시간	6	8	7	10	4
편차	-1	1	0	3	-3

따라서

$$(\text{분산})=\frac{1}{5}\{(-1)^2+1^2+0^2+3^2+(-3)^2\}=4$$

$$(\text{표준편차})=\sqrt{4}=2\ (\text{분})$$

답 분산: 4, 표준편차: 2분

▸분산에는 단위를 붙이지 않으며, 표준편차의 단위는 변량의 단위와 같다.

문제 1 다음은 무선 청소기 5종의 최대 사용 시간을 조사하여 나타낸 것이다. 최대 사용 시간의 분산과 표준편차를 각각 구하시오. (단, 표준편차는 반올림하여 일의 자리까지 구한다.)

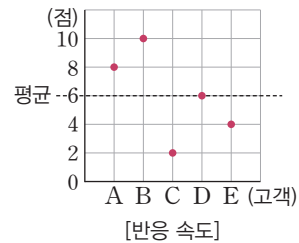
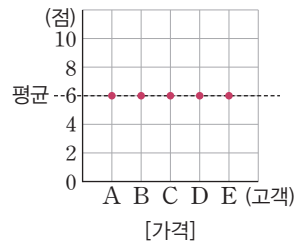
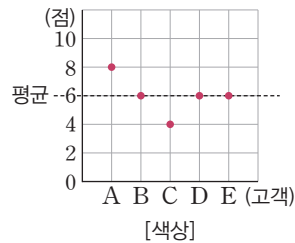
최대 사용 시간 (단위: 분)

36 46 26 52 35

문제 2 다음은 양궁 대표 선수 선발전에서 혜인이와 정미가 각각 화살을 10발씩 쏘아 얻은 점수를 조사하여 나타낸 것이다. 기록이 더 고른 사람을 대표 선수로 선발하려고 할 때, 혜인이와 정미 중에서 누구를 선발해야 하는지 말하시오.

양궁 점수 (단위: 점)

선수 \ 경기	1	2	3	4	5	6	7	8	9	10
혜인	8	3	9	10	6	10	10	4	6	4
정미	8	7	9	9	6	7	5	6	8	5

공학적 도구 활용하기

통그라미를 이용하여 대푯값과 산포도 구하기

공학적 도구를 사용하면 복잡한 자료의 대푯값과 산포도를 쉽게 구할 수 있다.

▸ 1 μm는 0.001 mm이다.

지름이 10 μm 이하인 미세 먼지를 PM10으로 나타내고, 지름이 2.5 μm 이하인 미세 먼지를 '초미세 먼지'라고 하며 PM2.5로 나타낸다.

다음은 2017년 4월 우리나라의 각 도시별 PM10의 대기 오염도 현황을 조사하여 나타낸 것이다.

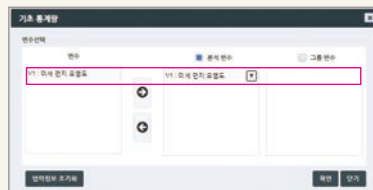
PM10의 대기 오염도 현황 (단위: $\mu g/m^3$)

51	54	56	71	53	59	51	59	60	68	57	59	63	59
78	57	68	63	67	61	70	59	64	52	65	74	60	61
73	65	65	68	67	61	53	73	61	56	54	55	55	65
49	49	59	49	49	48	56	53	56	62	59	69	73	58
78	52	58	60	61	58	40	46	56	50	50	43	56	51
37	46	52	51	52	54	61	40	59	52	42	65	44	

(출처: 국가통계포털, 2018)

통그라미를 이용하여 이 자료의 산포도를 구해 보자.

❶ 통그라미를 실행하고, 메뉴에서 '**편집>변수 설정**'을 클릭하여 변수명에 '미세 먼지 오염도'를 입력하고 저장한 후, V1열에 위의 자료의 변량을 세로로 차례대로 입력한다.

❷ 변량이 입력된 V1열을 선택한 후, 메뉴에서 '**통계>기초 통계량**'을 클릭하여 분석 변수에 'V1: 미세 먼지 오염도'를 추가하고 '**확인**'을 클릭하면 다음과 같이 PM10의 오염도에 대한 기초 통계량을 분석한 결과가 나타난다.

미세 먼지 오염도

분석변수	미세 먼지 오염도				
자료수	83	결측값수	0	합	4793.00
평균	57.75	중앙값	58.00	최빈값	59.00
최솟값	37.00	최댓값	78.00		
분산(n)	74.60	표준편차(n)	8.64		

탐구 통그라미 메뉴의 '파일>불러오기>예제 파일'에 제시된 자료 또는 국가통계포털의 자료 중의 하나를 불러와 대푯값 및 분산과 표준편차를 구해 보자.

스스로 확인하는 문제

1 대푯값

자료의 중심적인 경향이나 특징을 대표적으로 나타내는 값을 그 자료의 대푯값이라고 한다.

(1) **중앙값:** 자료의 변량을 작은 값부터 순서대로 나열할 때, 중앙에 위치하는 값

① 변량의 개수가 홀수이면 가운데 위치하는 값을 중앙값으로 한다.

② 변량의 개수가 짝수이면 가운데 위치하는 두 값의 평균을 중앙값으로 한다.

(2) **최빈값:** 자료의 변량 중에서 가장 많이 나타나는 값

예 ① 자료 '피자, 치킨, 떡볶이, 치킨'의 최빈값은 치킨이다.

② 자료 '1, 2, 1, 3, 4, 2, 5'의 최빈값은 1과 2이다.

2 산포도

자료의 분포 상태를 알아보기 위하여 변량들이 흩어져 있는 정도를 하나의 수로 나타낸 값을 그 자료의 산포도라고 한다.

(1) **편차:** 각 변량에서 평균을 뺀 값

➡ (편차)=(변량)−(평균)

(2) **분산과 표준편차**

① **분산:** 편차를 제곱한 값의 평균

➡ $(\text{분산})=\dfrac{\{(\text{편차})^2\text{의 총합}\}}{(\text{변량의 개수})}$

② **표준편차:** 분산의 양의 제곱근

➡ $(\text{표준편차})=\sqrt{(\text{분산})}$

③ 자료의 분산 또는 표준편차가 클수록 그 자료의 변량들이 평균을 중심으로 흩어져 있고, 작을수록 평균을 중심으로 모여 있다.

기본 문제

01 다음 자료의 중앙값을 구하시오.

(1) 1, 3, 7, 11, 15, 21, 28

(2) 5, −1, 2, −2, −3, 0, 4, 6

02 다음 자료의 최빈값을 구하시오.

(1) 4, 5, 5, 3, 2, 5

(2) 7, 5, 3, 2, 6, 4, 3

03 다음은 남학생 6명의 턱걸이 횟수를 조사하여 나타낸 것이다.

턱걸이 횟수 (단위: 회)

7	9	13	10	8	13

(1) 턱걸이 횟수의 평균을 구하시오.

(2) 각 변량의 편차를 구하시오.

(3) 턱걸이 횟수의 분산과 표준편차를 각각 구하시오.
(단, 표준편차는 반올림하여 소수점 아래 첫째 자리까지 구한다.)

04 다음은 한글 단어 퀴즈 대회에 참여한 외국인 15명이 맞힌 단어의 개수를 조사하여 나타낸 것이다. 맞힌 단어의 개수의 평균, 중앙값, 최빈값을 각각 구하시오.

맞힌 단어의 개수 (단위: 개)

25	18	34	45	31	25	27	25
20	27	21	30	19	25	24	

05 오른쪽 줄기와 잎 그림은 어느 제과점에서 생산하는 15종류의 빵에 대하여 하루 동안의 판매량을 조사하여 나타낸 것이다. 빵의 판매량의 중앙값과 최빈값을 각각 구하시오.

빵의 판매량 (1|0은 10개)

줄기	잎
1	0 2 7
2	2 3 4 4 5 6 8
3	1 4 4 8 9

06 자료 'a, 8, 9, 13, 17, 12'의 중앙값이 11일 때, a의 값을 구하시오.

07 다음은 어느 병원에서 환자 5명의 접수 후 대기 시간에 대한 편차를 조사하여 나타낸 것이다. 대기 시간의 평균이 12분일 때, 환자 B의 대기 시간을 구하시오.

대기 시간 (단위: 분)

환자	A	B	C	D	E
편차	8		2	-6	-1

08 어떤 자료의 각 변량에 대한 편차가 다음과 같다.

a	-3	5	-4	3	-1

(1) a의 값을 구하시오.

(2) 이 자료의 분산과 표준편차를 각각 구하시오.

09 다음은 어느 축구 경기에서 5개의 팀이 기록한 득점과 실점을 조사하여 나타낸 것이다.

득점과 실점 (단위: 점)

팀	A	B	C	D	E
득점	1	1	5	2	1
실점	2	1	2	3	2

(1) 득점과 실점의 표준편차를 각각 구하시오.
(단, 표준편차는 반올림하여 소수점 아래 첫째 자리까지 구한다.)

(2) 평균을 대푯값으로 할 때, 득점과 실점의 분포 중에서 산포도가 더 큰 것을 말하시오.

10 문제 해결

다음 두 조건을 모두 만족시키는 a와 b의 값을 각각 구하시오.

> (가) 5, 8, 13, 15, a의 중앙값은 8이다.
> (나) 2, 15, a, b, 14의 중앙값은 12이고, 평균은 $b-2$이다.

11 다음 자료의 분산이 4일 때, 양수 a의 값을 구하시오.

> a　　4　　$2a+4$　　$a+4$　　$a+3$

12 자료 'a, b, c'의 평균이 10이고 분산이 6일 때, 자료 '9, a, b, c, 11'의 표준편차를 구하시오.

2 상관관계

강수량이 증가할 때 우산 판매량도 증가한다고 합니다. 비가 자주 오게 되면 우산을 쓸 기회가 많아지기 때문에, 강수량의 증가가 우산 판매량의 증가에 직접적으로 영향을 끼쳐서 강수량과 우산 판매량이 함께 증가하는 관계가 있음을 알 수 있습니다.

한편, 아이스크림 판매량이 증가할 때 선글라스 판매량도 증가한다고 합니다. 날씨가 더워지면 아이스크림을 더 많이 사 먹게 될 뿐만 아니라 자외선으로부터 눈을 보호하기 위해 평소보다 선글라스를 더 많이 사게 되기 때문입니다. 아이스크림 판매량과 선글라스 판매량의 관계는 날씨에 의한 것이지 서로 직접적인 영향을 끼치는 것은 아닙니다. 그러나 이 둘 사이에는 함께 증가하거나 함께 감소하는 관계가 있음을 알 수 있습니다.

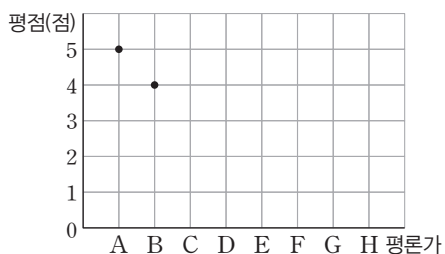
이 단원에서는 자료를 산점도로 나타내는 방법과 상관관계에 대하여 알아봅니다.

• 자료를 그래프로 나타내기

1 다음은 영화 평론가 8명이 어떤 영화에 대하여 매긴 평점을 조사하여 나타낸 것이다.

영화 평점 (단위: 점)

평론가	A	B	C	D	E	F	G	H
평점	5	4	3	2	4	4	3	3

위의 자료를 이용하여 오른쪽 그래프를 완성하시오.

산점도와 상관관계

학습 목표 • 자료를 산점도로 나타내고, 이를 이용하여 상관관계를 말할 수 있다.

다가서기

산점도와 상관관계란 무엇인가?

다음은 북반구에 있는 여러 도시의 위도와 1971년부터 2000년까지 7월 1일의 평균 기온을 조사하여 나타낸 것이다.

평균 기온

도시	위도(°)	기온(℃)	도시	위도(°)	기온(℃)	도시	위도(°)	기온(℃)
더블린	53.3	14.2	서울	37.6	23.1	카이로	30	27.8
로마	41.9	22.8	시카고	41.9	21.4	콜카타	22.6	28.8
모스크바	55.8	18.6	아테네	38	26.4	파리	48.9	18.7
방콕	13.8	29.2	오슬로	59.9	14.6	헬싱키	60.2	15.7
베이징	39.9	25.4	이스탄불	41	23.2	홍콩	22.4	29.4

(출처: 기상자료개방포털, 2018)

1. 위도가 북위 50° 이상인 도시를 모두 찾고, 각 도시의 평균 기온을 말해 보자.
2. 위도가 북위 30° 미만인 도시를 모두 찾고, 각 도시의 평균 기온을 말해 보자.
3. 위도와 평균 기온 사이에 어떤 관계가 있는지 말해 보자.

위의 생각 열기에서 위도가 높은 도시의 평균 기온이 대체로 낮고, 위도가 낮은 도시의 평균 기온은 대체로 높음을 알 수 있다.

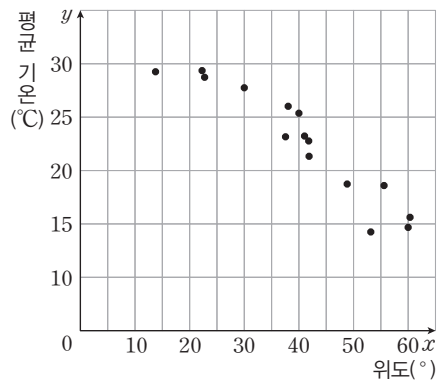
이와 같이 우리 생활 주변에서 얻는 자료의 두 변량 중에는 서로 관계가 있는 것도 있고 전혀 관계가 없는 것도 있다. 이때 두 변량을 그래프로 나타내면 이들 사이의 관계를 알아보기 편리한 경우가 있다.

앞의 생각 열기의 자료를 이용하여 북반구 도시의 위도와 평균 기온 사이의 관계를 그래프로 나타내 보자.

각 도시의 위도를 북위 $x°$, 평균 기온을 y ℃라고 할 때, 두 변량 x, y의 순서쌍 (x, y)를 좌표평면 위에 점으로 나타내면 [그림 1]과 같은 그래프로 나타낼 수 있다.

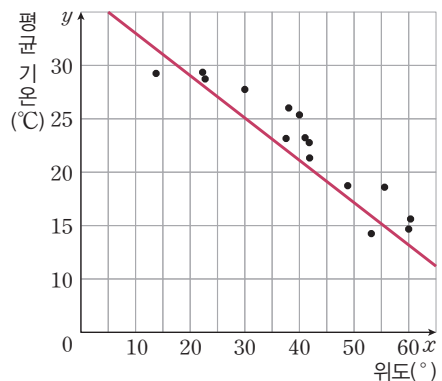
[그림 1]

평균 기온(℃) y: 0, 10, 15, 20, 25, 30
위도(°) x: 10, 20, 30, 40, 50, 60

이와 같이 어떤 자료에서 두 변량 x와 y에 대하여 순서쌍 (x, y)를 좌표평면 위에 점으로 나타낸 그래프를 x와 y의 **산점도**라고 한다. 이를테면 [그림 1]은 북반구 도시의 위도와 평균 기온의 산점도이다.

▸ 산점도(散點圖)에서 산(散)은 흩어진다, 점(點)은 점, 도(圖)는 그림을 나타내는 한자이다.

이 산점도에서 점들은 어느 정도 흩어져 있지만, [그림 2]와 같이 대체로 오른쪽 아래로 향하는 한 직선을 중심으로 그 주위에 가까이 분포되어 있다고 볼 수 있다.

[그림 2]

평균 기온(℃) y: 0, 10, 15, 20, 25, 30
위도(°) x: 10, 20, 30, 40, 50, 60

따라서 위도가 높아질수록 대체로 평균 기온이 낮아지고, 위도가 낮아질수록 대체로 평균 기온이 높아짐을 알 수 있다.

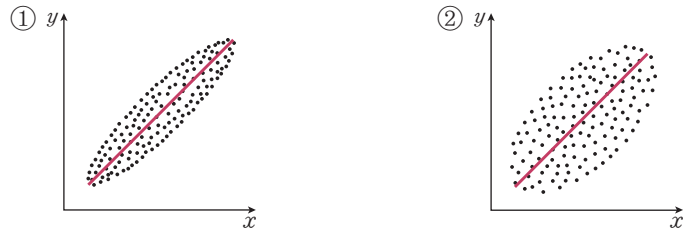
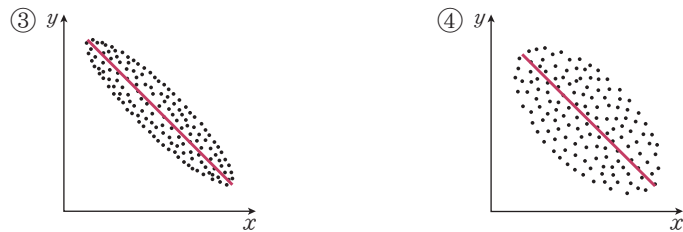
이 산점도로부터 위도와 평균 기온 사이에 어떤 관계가 있음을 알 수 있는데, 이와 같은 두 변량 사이의 관계를 **상관관계**라고 한다.

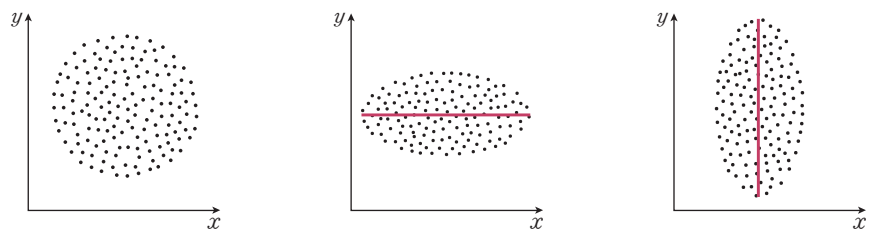
▸ 핼리(Halley, E., 1656~1742)
영국의 천문학자로, 고도와 기압의 관측 자료를 좌표평면에 나타낸 것이 산점도의 기원이 되었다고 한다.

어떤 자료에서 두 변량 x와 y에 대하여 x의 값이 커짐에 따라 y의 값도 대체로 커지는 관계가 있을 때, 두 변량 사이에는 '양의 상관관계가 있다'고 한다. 또 x의 값이 커짐에 따라 y의 값이 대체로 작아지는 관계가 있을 때, 두 변량 사이에는 '음의 상관관계가 있다'고 한다.

이를테면 위의 자료에서 위도와 평균 기온 사이에는 음의 상관관계가 있음을 알 수 있다.

양 또는 음의 상관관계가 있는 산점도에서 점들이 한 직선에 가까이 분포되어 있을수록 '상관관계가 강하다'고 하고, 흩어져 있을수록 '상관관계가 약하다'고 한다.

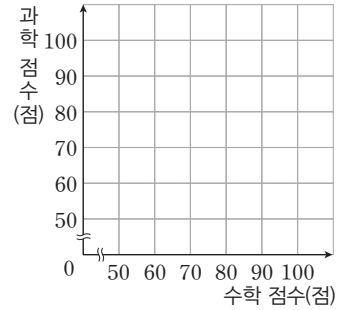
다음 산점도에서 ①은 ②보다 양의 상관관계가 강한 경우이고, ②는 ①보다 양의 상관관계가 약한 경우이다.

다음 산점도에서 ③은 ④보다 음의 상관관계가 강한 경우이고, ④는 ③보다 음의 상관관계가 약한 경우이다.

한편, 어떤 자료에서 두 변량 x와 y에 대하여 x의 값이 커짐에 따라 y의 값이 커지는지 또는 작아지는지 그 관계가 분명하지 않은 경우에, 두 변량 사이에는 '상관관계가 없다'고 한다.

예를 들어 다음 그림과 같이 산점도에서 점들이 한 직선에 가까이 분포되어 있지 않거나, x축 또는 y축에 평행한 직선에 가까이 분포되어 있는 경우는 두 변량 x와 y 사이에 상관관계가 없다.

문제 1 다음은 어느 반 학생 20명의 수학과 과학 점수를 조사하여 나타낸 것이다.

수학과 과학 점수 (단위: 점)

번호	수학	과학	번호	수학	과학	번호	수학	과학	번호	수학	과학
1	90	86	6	92	64	11	73	84	16	92	95
2	64	65	7	84	98	12	82	83	17	65	70
3	94	89	8	72	85	13	78	59	18	76	72
4	57	62	9	52	62	14	68	64	19	83	88
5	82	74	10	86	84	15	54	55	20	95	93

(1) 수학과 과학 점수의 산점도를 오른쪽 좌표평면 위에 그리시오.

(2) 수학과 과학 점수 사이에 어떤 상관관계가 있는지 말하시오.

공학적 도구를 이용하여 산점도와 상관관계를 어떻게 알아보는가?

통그라미를 이용하여 산점도를 그리는 방법을 알아보자.

▸ 통그라미 누리집(http://tong.kostat.go.kr)에서 통그라미를 실행할 수 있다.

오른쪽 표는 어느 날 우리나라 여러 지점의 풍속과 평균 파고를 조사하여 나타낸 것이다.

1 통그라미를 실행하여 '풍속'과 '평균 파고'의 변량을 각각 차례대로 입력해 보자.

2 메뉴에서 '그래프>산점도'를 클릭하고, 가로축 변수에 '풍속', 세로축 변수에 '평균 파고'를 선택한 후 '확인'을 클릭하여 풍속과 평균 파고의 산점도를 그려 보자.

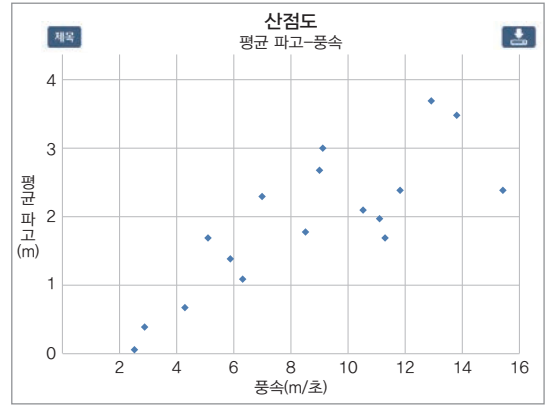
풍속과 평균 파고

지점	풍속(m/초)	평균 파고(m)
거문도	11.1	2.0
거제도	15.4	2.4
덕적도	2.9	0.4
동해	7.0	2.3
마라도	10.5	2.1
부안	5.9	1.4
서귀포	13.8	3.5
신안	2.3	0.0
외연도	6.3	1.1
울릉도	9.0	2.7
울산	12.9	3.7
울진	9.1	3.0
인천	4.3	0.7
추자도	8.5	1.8
칠발도	5.1	1.7
통영	11.3	1.7
포항	11.8	2.4

(출처: 날씨누리, 2018)

앞의 함께하기에서 통그라미를 이용하여 그린 풍속과 평균 파고의 산점도는 오른쪽과 같다.

이 산점도로부터 풍속과 평균 파고 사이에 양의 상관관계가 있음을 알 수 있다.

탐구 문제 2

▸날씨누리(http://www.weather.go.kr)에서 여러 가지 날씨 자료를 조사할 수 있다.

우리 생활 주변에서 두 변량이 있는 실제 자료를 조사하여 다음에 답하시오.

(1) 조사한 자료에서 비교할 두 변량을 정하고, 통그라미를 이용하여 산점도를 그리시오.

(2) 조사한 자료의 두 변량 사이에 어떤 상관관계가 있는지 말하시오.

생각이 크는 수학

오른쪽 산점도는 피자 14종의 150 g당 포화 지방의 양(g)과 콜레스테롤의 양(mg)을 조사하여 나타낸 것이다.

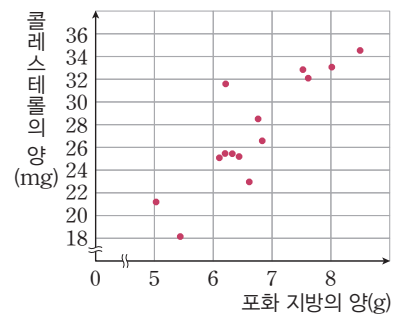
(출처: 한국소비자원, 「피자 품질 시험 결과」)

1 포화 지방의 양과 콜레스테롤의 양 사이에 어떤 상관관계가 있는지 말해 보자.

2 포화 지방의 양이 7 g 미만인 피자의 비율을 구해 보자.

3 포화 지방의 양이 7 g 이상인 피자 중에서 콜레스테롤의 양이 30 mg 이상인 피자의 비율을 구해 보자.

알콩 달콩 수학+ 활동

의사소통 태도 및 실천

통계 포스터 만들기

우리가 생활에서 쉽게 접할 수 있는 두 변량의 상관관계에 대하여 궁금해하던 주제를 하나 정하고, 그 주제에 대한 자료를 수집하여 서로 비교해 보는 프로젝트를 다음과 같이 수행해 보자. 또 프로젝트를 수행한 결과를 포스터로 만들어 보자.

통계 포스터의 작성 방법은 다음과 같다.

주제 정하기 → 자료 수집 방법 정하기 → 자료 수집 → 표와 그래프 만들기 → 해석하기 → 포스터 만들기

[예시 주제]

- 우리 반 학생들의 통학 거리와 등교 소요 시간
- 우리 반 학생들의 각 가정에서 지출하는 전기 요금과 수도 요금
- 한 햄버거 가게에서 파는 햄버거의 종류별 열량과 함유 나트륨의 양

활동 1 모둠을 정하여 모둠별로 두 변량의 상관관계를 비교할 주제와 그것을 선정한 이유, 그리고 양의 상관관계나 음의 상관관계 중에서 어느 쪽으로 예측되는지를 다음에 적어 보자.

[1] 조사할 주제:

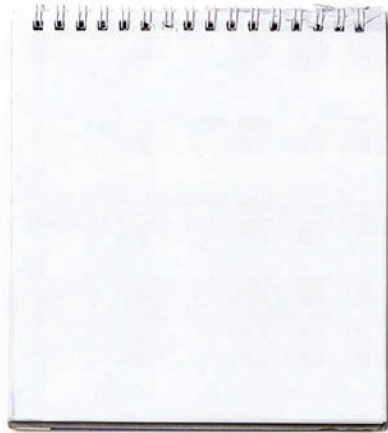
[2] 주제를 선정한 이유:

[3] 예측되는 결과:

활동 2 모둠별로 조사할 주제에 대하여 자료를 수집할 방법과 정리 방법을 토의하여 정하고, 다음에 적어 보자.

[1] 자료 수집 방법:

[2] 정리 방법:

활동 3 모둠에서 토의한 방법에 따라 각자의 역할을 정하여 필요한 것을 조사하여 정리하고, 이를 표와 산점도로 각각 나타내 보자.

[표]

[산점도]

활동 4 산점도를 보고 두 변량 사이에 양의 상관관계가 있는지, 음의 상관관계가 있는지, 상관관계가 없는지에 대하여 토의하여 처음 예측한 것이 맞는지 그 결과를 다음에 적어 보자.

활동 5 주제의 선정 동기, 자료 수집 방법, 정리 방법과 산점도, 분석의 결과로부터 알게 된 사항을 포함하는 포스터를 만들고 이를 모둠별로 발표해 보자.

활동 6 다음 평가표의 해당하는 곳에 ○표를 하고, '통계 포스터 만들기'를 수행하며 배운 점과 아쉬운 점을 이야기해 보자.

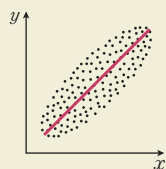
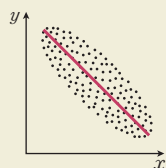
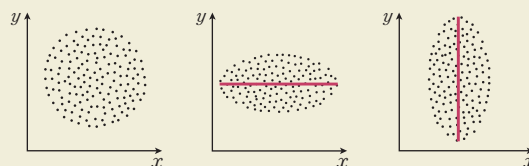
평가 요소	우수	보통	미흡
알고 싶었거나 궁금했던 것이 해결되었는가?			
주제에 맞는 계획을 세우고 타당한 자료 수집 방법을 선택하였는가?			
결과를 해석하는 과정에서 표나 산점도 등을 올바르게 사용하였는가?			

스스로 확인하는 문제

1 산점도와 상관관계

(1) **산점도:** 어떤 자료에서 두 변량 x와 y에 대하여 순서쌍 (x, y)를 좌표평면 위에 점으로 나타낸 그래프

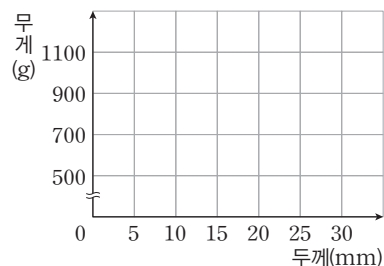
(2) **상관관계**

① **양의 상관관계:** 두 변량 x와 y에 대하여 x의 값이 커짐에 따라 y의 값도 대체로 커지는 관계

② **음의 상관관계:** 두 변량 x와 y에 대하여 x의 값이 커짐에 따라 y의 값이 대체로 작아지는 관계

③ 두 변량 x와 y에 대하여 x의 값이 커짐에 따라 y의 값이 커지는지 또는 작아지는지 그 관계가 분명하지 않은 경우에, 두 변량 사이에는 '상관관계가 없다'고 한다.

(3) 양 또는 음의 상관관계가 있는 산점도에서 점들이 한 직선에 가까이 분포되어 있을수록 '상관관계가 강하다'고 하고, 흩어져 있을수록 '상관관계가 약하다'고 한다.

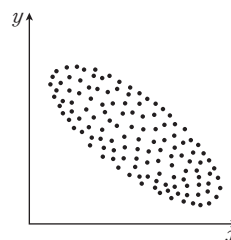
기본 문제

01 다음은 책 10권의 두께와 무게를 조사하여 나타낸 것이다. 이 자료에서 책의 두께와 무게의 산점도를 오른쪽 좌표평면 위에 그리시오.

책의 두께와 무게

두께(mm)	무게(g)	두께(mm)	무게(g)
9	710	29	990
12	650	20	860
32	1230	24	920
10	590	18	970
13	720	16	800

02 다음 보기 중에서 두 변량 사이의 산점도가 대체로 오른쪽 그림과 같은 모양이 되는 것을 모두 고르시오.

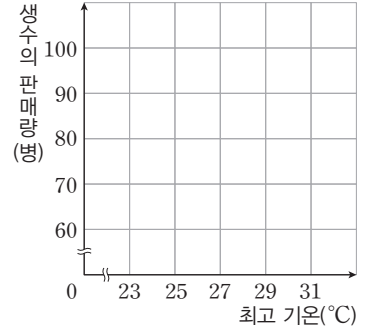
보기
ㄱ. 어떤 물건의 가격과 판매량
ㄴ. 여름철 기온과 냉방비
ㄷ. 신발 크기와 그 신발의 가격
ㄹ. 자동차의 이동 거리와 남은 기름의 양

표준 문제

03 다음은 어느 편의점에서 그날의 최고 기온과 하루 동안 판매된 500 mL짜리 생수의 개수를 10일 동안 조사하여 나타낸 것이다.

최고 기온과 생수의 판매량

최고 기온(℃)	31	28	30	26	25	30	29	23	28	27
생수의 판매량(병)	95	80	92	67	72	96	94	67	78	75

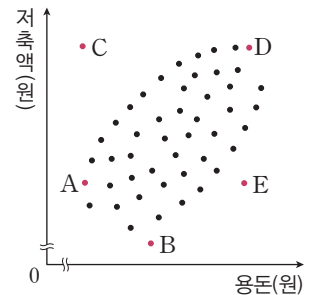
(1) 최고 기온과 생수의 판매량의 산점도를 오른쪽 좌표평면 위에 그리시오.

(2) 최고 기온과 생수의 판매량 사이에 어떤 상관관계가 있는지 말하시오.

04 오른쪽 산점도는 학생 50명의 용돈과 저축액을 조사하여 나타낸 것이다.

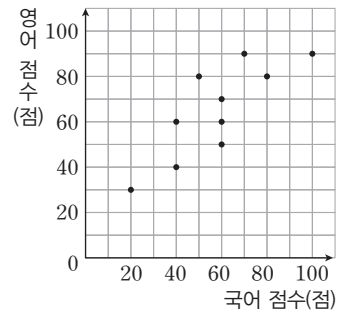
(1) 용돈과 저축액 사이에 어떤 상관관계가 있는지 말하시오.

(2) A, B, C, D, E에 해당하는 학생 중에서 특별히 용돈에 비하여 저축액이 많은 학생은 누구인지 말하시오.

발전 문제

05 문제 해결

오른쪽 산점도는 민수네 반 학생 10명의 국어 점수와 영어 점수를 조사하여 나타낸 것이다.

(1) 국어 점수가 영어 점수보다 높은 학생의 비율을 구하시오.

(2) 국어 점수가 60점 이상인 학생 중에서 영어 점수가 80점 이상인 학생의 비율을 구하시오.

단원을 마무리하는 문제

01 어떤 자료의 변량을 작은 값부터 순서대로 나열하면 '3, 6, 8, 10, x'이다. 이 자료의 평균과 중앙값이 같을 때, x의 값은?

① 10 ② 11 ③ 12
④ 13 ⑤ 14

02 다음 자료의 최빈값은?

5 9 11 7 13 8 5 9 5 11 8 5

① 5 ② 7 ③ 8
④ 9 ⑤ 11

03 다음 자료 중에서 중앙값이 평균보다 자료의 중심적인 경향을 더 잘 나타내는 것은?

① 3, 4, 4, 6, 8, 9, 9, 10, 11
② 1, 2, 5, 10, 20, 26, 32, 38, 46
③ 0.3, 0.1, 1.6, 1, 0.7, 0.4, 0.2, 0.5, 1
④ 32, 2, 4, 1, 5, 2, 3, 4
⑤ -2, -1, 5, 8, -3, 0, 2, -10, 1

04 아래 표는 두 전철역 A와 B에서 각각 5명의 승객을 대상으로 전철을 기다린 시간을 조사하여 나타낸 것이다. 다음 중에서 옳지 않은 것은?

전철을 기다린 시간 (단위: 분)

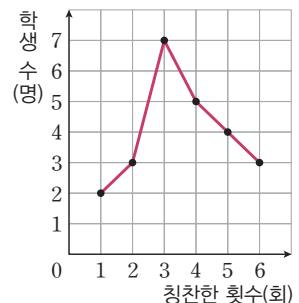
A 역	8	6	4	6	6
B 역	10	4	5	7	4

① A 역의 자료의 중앙값은 최빈값과 같다.
② B 역의 자료의 평균은 중앙값보다 크다.
③ B 역의 자료의 최빈값은 1개이다.
④ A 역과 B 역의 자료의 평균은 서로 같다.
⑤ A 역과 B 역의 자료의 중앙값은 서로 같다.

05 다음 자료의 중앙값은 9이고 최빈값은 7일 때, a의 값은?

$a+4$ 9 7 7 10 9 $3a+1$

① 2 ② 3 ③ 4
④ 5 ⑤ 6

06 오른쪽 꺾은선 그래프는 학생 24명을 대상으로 하루 동안 다른 사람을 칭찬한 횟수를 조사하여 나타낸 것이다. 칭찬한 횟수의 중앙값을 a회, 최빈값을 b회라고 할 때, $a+b$의 값을 구하시오.

07 다음 자료 중에서 변량 6의 편차는?

10　9　8　6　11　7　5

① -2　② -1　③ 0
④ 1　⑤ 2

08 다음 중에서 자료 '4, 5, 6, 6, 7, 7, 7'에 대한 설명으로 옳지 않은 것을 모두 고르면? (정답 2개)

① 이 자료의 중앙값은 6이다.
② 이 자료의 최빈값은 7이다.
③ 이 자료의 평균은 6.5이다.
④ 이 자료의 편차의 합은 0이다.
⑤ 이 자료의 표준편차는 1이다.

09 자료 '$a-1$, $a+2$, $a+3$, $a+4$'의 분산은?

① 2　② $\frac{5}{2}$　③ 3
④ $\frac{7}{2}$　⑤ 4

10 다음 자료의 표준편차는?

7　4　3　6　5

① $\sqrt{2}$　② $\frac{\sqrt{2}}{2}$　③ $\sqrt{3}$
④ $\frac{\sqrt{3}}{2}$　⑤ 2

11 자료 'a, b, c'의 평균은 0, 분산은 1이고 자료 'd, e, f'의 평균은 0, 분산은 2이다. 이때 자료 'a, b, c, d, e, f'의 분산을 구하시오.

12 자료 '2, a, 4, b, 6'의 평균이 4, 표준편차가 $\sqrt{2}$일 때, a^2+b^2의 값은?

① 22　② 26　③ 29
④ 32　⑤ 34

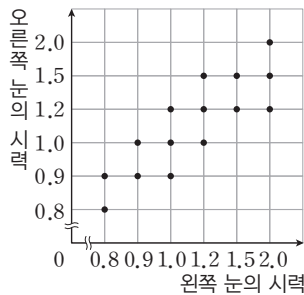
13 다음은 어느 식당의 고객 10명에게 식당의 청결도에 대한 점수를 조사하여 나타낸 것이다. 청결도 점수의 표준편차를 구하시오.

청결도 점수

점수(점)	1	2	3	4	5
고객(명)	1	2	3	4	0

14 다음 중에서 산점도에 대한 설명으로 옳은 것은?

① 산점도는 산포도를 그래프로 나타낸 것이다.
② 두 변량의 평균이 어떤 관계를 가지는지 산점도로 확인할 수 있다.
③ 점들이 한 직선에 가까이 분포되어 있는 산점도는 양의 상관관계를 나타낸다.
④ 음의 상관관계를 나타내는 산점도는 점들이 기울기가 음수인 한 직선에 가까이 분포되어 있다.
⑤ 점들이 한 직선에 가까이 분포된 산점도는 상관관계가 큰 것을 나타낸다.

15 오른쪽 산점도는 중학교 학생 15명의 좌우 시력을 조사하여 나타낸 것이다. 왼쪽 눈의 시력이 오른쪽 눈의 시력보다 좋은 학생의 비율은?

① $\frac{1}{3}$ ② $\frac{2}{3}$ ③ $\frac{1}{5}$
④ $\frac{2}{5}$ ⑤ $\frac{3}{5}$

[16~19] 서술형 풀이 과정과 답을 써 보자.

16 다음은 학생 10명의 제기차기 횟수를 조사하여 나타낸 것이다. 평균, 중앙값, 최빈값을 각각 구하고, 그중에서 대푯값으로 가장 적절한 것을 말하시오.

제기차기 횟수 (단위: 회)

10 8 11 5 11 7 5 8 7 78

17 양수 a에 대하여 다음 자료의 분산이 3.2일 때, 평균을 구하시오.

$2 \quad 2a \quad a-1 \quad a+4 \quad a$

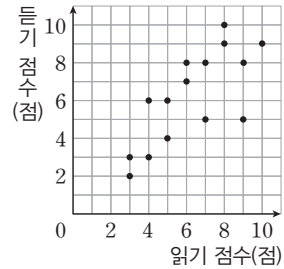
18 다음은 두 축구팀 A와 B의 월별 득점을 조사하여 나타낸 것이다.

월별 득점 (단위: 점)

팀 \ 월	3	4	5	6	7	8
A	8	10	9	7	13	13
B	9	9	10	12	11	9

(1) 두 팀의 득점의 분산을 각각 구하시오.

(2) 두 팀 중에서 월별 득점이 더 고른 팀을 말하시오.

19 다음 산점도는 읽기와 듣기를 각각 10점 만점으로 평가하는 영어 능력 시험에서 응시자 15명의 점수를 조사하여 나타낸 것이다.

(1) 듣기 점수의 최빈값을 구하시오.

(2) 읽기 점수와 듣기 점수가 모두 7점 이상인 사람을 합격시킨다고 할 때, 전체 응시자 중에서 합격자의 비율을 구하시오.

자기 평가 정답을 맞힌 문항에 ○표를 하고 결과를 점검한 다음, 이 단원의 학습 목표를 얼마나 성취했는지 스스로 평가하고, 학습 보충 계획을 세워 보자.

문항 번호	학습 목표	성취도
01 02 03 04 05 06 16	중앙값, 최빈값, 평균의 의미를 이해하고, 이를 구할 수 있는가?	☺ 😐 ☹
07 08 09 10 11 12 13 17 18	분산과 표준편차의 의미를 이해하고, 이를 구할 수 있는가?	☺ 😐 ☹
14 15 19	자료를 산점도로 나타내고, 이를 이용하여 상관관계를 말할 수 있는가?	☺ 😐 ☹

0개~10개 개념 학습이 필요해요! 11개~13개 부족한 부분을 검토해 봅시다! 14개~16개 실수를 줄여 봅시다! 17개~19개 훌륭합니다!

학습 보충 계획:

운동선수의 기록과 평균

운동선수의 능력을 평가하거나 경기에서 순위를 정할 때, 수영이나 달리기 같은 경우에는 선수의 기록이 작은 값일수록 좋고, 양궁이나 사격 같은 경우에는 선수의 기록이 큰 값일수록 좋다. 한편, 리듬 체조는 여러 명의 심판이 매긴 점수의 평균으로 기록을 정하기도 한다.

여기서는 앞에서 배운 평균과 편차가 운동선수의 기록 향상에 어떤 의미를 갖는지 알아보기로 한다.

1 수영, 달리기

수영이나 달리기 경기에서는 기록의 측정값이 작을수록 좋기 때문에 선수들의 기록 중에서 가장 작은 값이 의미를 갖는다.

선수들은 자신의 기록을 단축하기 위해 많은 훈련을 하게 되는데, 이런 종목에서는 기록 측정값의 평균과 편차를 낮출수록 다음 경기에서 더 좋은 기록을 세울 가능성이 커진다.

2 양궁, 사격

양궁이나 사격 경기에서는 주어진 시간 내에 일정한 횟수를 쏘아서 얻은 점수의 합으로 순위를 매긴다. 따라서 이런 경기의 훈련에서는 한 선수의 점수의 합 중에서 가장 큰 값이 의미를 갖는다.

자신의 기록을 향상하기 위해서는 점수의 편차는 작아지고 평균은 높아지도록 훈련할 필요가 있다.

3 리듬 체조

리듬 체조 경기는 4~6명의 심판이 채점을 한 후에 최고점과 최하점을 하나씩 제외한 2~4명의 점수의 평균으로 점수를 매기는데, 이것은 채점 자료의 이상점을 제거하고 구한 절사 평균이다(211쪽 참조).

체조나 다이빙, 피겨 스케이팅 등의 경기에서는 여러 심판이 매기는 점수의 편차가 너무 크게 나타나는 경우를 방지하기 위하여 이와 같은 절사 평균을 이용한다.

색깔로 보는 상관관계

통계적 분석이 필요한 자료의 변량을 특성에 따라 여러 가지 색깔로 나타낸 것을 '히트 맵(heat map)'이라고 하는데, 이 방법은 100여 년 전부터 통계 처리에 이용되었다고 한다.

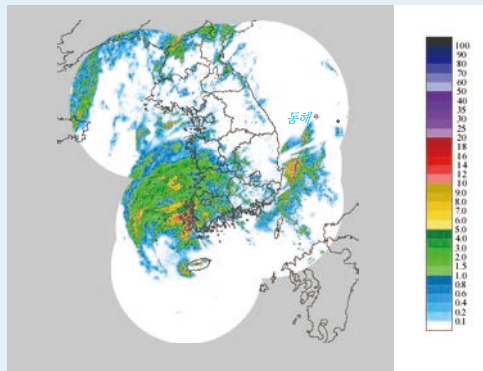
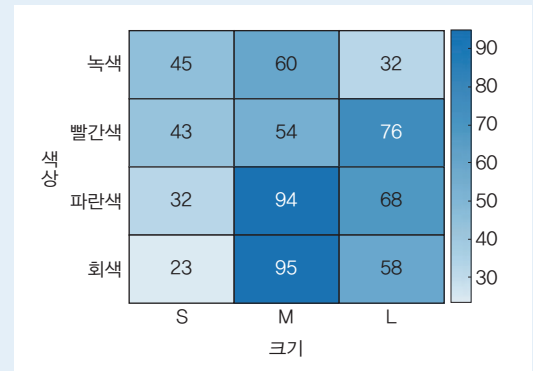
히트 맵은 많은 양의 데이터를 시각화하는 데 편리하기 때문에 컴퓨터 성능의 발전에 힘입어 최근 들어 폭넓게 이용되고 있다.

또한 어떤 자료에서 변량 사이의 관계를 직사각형의 격자 모양으로 배열한 후에 비슷한 관계를 유사한 색 영역으로 표시할 수 있어서 두 변량의 상관관계를 비교하는 데 편리하다.

기상 상태를 알려 주는 레이더 영상도 히트 맵의 일종이다. [그림 1]은 태풍이 지나갈 때의 강수 구역을 나타내고 있는데, 붉은색일수록 비가 많이 내리는 지역을 나타낸다. [그림 2]는 어느 티셔츠 공장에서 하루 동안 들어온 색상과 크기별 주문량을 히트 맵으로 나타낸 것인데, 색깔이 짙을수록 주문량이 많음을 나타낸다. 이로부터 중간 크기의 회색 티셔츠의 주문이 가장 많았음을 알 수 있다.

[그림 1] 강수 구역의 레이더 영상

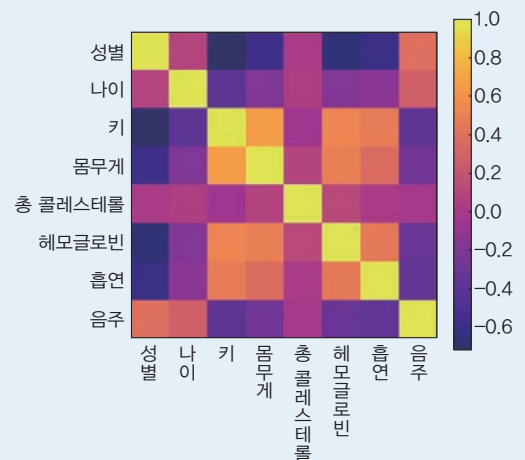
색상 \ 크기	S	M	L
녹색	45	60	32
빨간색	43	54	76
파란색	32	94	68
회색	23	95	58

(색상 막대: 90, 80, 70, 60, 50, 40, 30)

[그림 2] 티셔츠 주문량

오른쪽 그림은 2016년 건강 검진 수검자 100만 명의 검진 내역과 문진 항목을 바탕으로 만든 히트 맵인데, 두 항목 사이의 상관관계는 색깔이 밝을수록 강하고 어두울수록 약함을 오른쪽 세로 막대의 수로 나타내고 있다.

이 그림으로부터 헤모글로빈의 수치는 키, 몸무게, 흡연, 총 콜레스테롤의 순으로 관련성이 있음을 알 수 있다.

(출처: Wilkinson, L., Friendly, M., 「The History of the Cluster Heat Map」)

방대한 자료를 효과적으로 분석하는 방법은 무엇일까요?

빅 데이터(big data) 전문가는 엄청난 양의 통계 자료에서 의미 있는 분석 결과를 만들어 내는 일을 합니다. 각종 인터넷 등에서 실시간으로 쏟아지는 방대한 자료를 '빅 데이터'라고 하는데, 빅 데이터 전문가는 빅 데이터를 어디에, 어떻게 활용할 것인지 기획하는 일에서부터, 분석할 빅 데이터의 자원을 찾고, 프로그램을 만든 뒤 통계적으로 분석하는 작업을 합니다. 또 그 결과물을 다양한 분야에서 활용하는 역할도 합니다.

'빅 데이터'라는 용어는 미국 실리콘 밸리에서 처음 등장했습니다. 2000년대 초반 전 세계적으로 인터넷이 보급되면서 미국의 신흥 정보 서비스 기업들이 폭발적인 성장을 하게 되었습니다. 이에 따라 인터넷상의 정보의 양도 가늠할 수 없을 정도로 늘어났습니다. 인터넷뿐만 아니라 스마트폰이 보급되면서부터는 소셜 네트워크 서비스(SNS) 등을 통해 개개인이 언제든 간편하게 정보의 생산과 소비를 할 수 있게 되었습니다. 이러한 개인의 정보에 기업들의 관심이 집중되면서 자연스럽게 빅 데이터 전문가에 대한 수요가 크게 늘어났습니다.

이제는 빅 데이터가 우리 생활에 끼치는 영향을 쉽게 찾아볼 수 있습니다. 예를 들어 대통령이나 국회의원 선거 당일 실시간 출구 조사, 인터넷 서점에서 고객의 독서 성향을 파악해서 고객의 취향에 맞는 책을 추천하는 서비스, 특정한 전염병의 확산 경로와 속도의 추정, 태풍의 이동 경로 예측, 특정한 운동 선수의 공격 또는 수비 형태의 특징 파악 등은 모두 빅 데이터 분석의 결과로부터 이루어집니다.

빅 데이터 전문가가 되기 위해서는 빅 데이터를 활용하는 데 필요한 통계학, 수학, 컴퓨터 공학, 산업 공학뿐만 아니라 경제학과 심리학 등의 지식과 기술이 필요합니다.

(출처: 워크넷, 2018 / 연합마이더스, 2019)